The Collagens of the Ctenophore *Pleurobrachia bachei*

Nathan Churches[1,2], Billie J. Swalla[1,2], Andrea Kohn[1,3], Leonid Moroz[1,3]

FHL/REU/Blinks/BEACON
Research Experience For Undergrads
Summer, 2012

[1]University of Washington, Friday Harbor Laboratories, Friday Harbor, WA 98250
[2]Department of Biology, University of Washington, Seattle, WA 98195
[3]Department of Biology, The Whitney Laboratory for Marine Bioscience, St. Augustine, FL 32080

Contact Information:
Nathan Churches
nthnchrchs@yahoo.com

(360) 870 8589

The Collagens of the Ctenophore *Pleurobrachia bachei*

Nathan Churches

ABSTRACT

The evolution of multicellular animals required the development of epithelial tissues that function in controlling the transport of molecules from one environment to another. Collagen proteins are crucial to the formation of epithelial tissues, and are therefore critical in understanding the origins of multicellularity and Metazoa. We looked for collagen proteins in the recently sequenced Ctenophore *Pleurobrachia bachei* genome, using collagen sequences from both *Homo sapiens* and basal organisms as templates. We discovered that *P. bachei* has 7 distinct type IV collagen sequences along three genomic scaffolds. Two of these were aligned in an in-line pattern along the genome, while another two were aligned in a head-to-head fashion, indicating both traditional and inverted gene duplication events. Furthermore, collagen IV intra-genic investigations showed that *P. bachei* has a more diverse set of α-chains than *Homo sapiens.* In addition, we found 4 conserved collagen domain proteins which we were unable to assign to specific collagen families. Given recent evidence suggesting that Ctenophores are the most basal phyla of the Metazoans, our findings suggest that the common ancestor to all Metazoa contained a much more developed collagen profile than previously appreciated.

INTRODUCTION

The advent of multicellularity involved the evolution of epithelial tissues that were capable of controlling the transfer of molecules from one environment to another. This task is carried out by epithelia in modern multicellular animals. Presumably, a sort of 'proto-epithelia' evolved in early colonial animals, where dividing cells remained adhered to one another and were able to control their internal environments, thus forming distinguished multicellular organisms (Leys and Reisgo, 2011). Collagens provide the structural matrix of epithelial tissues, playing a crucial role in animal morphology and physiology. Therefore, understanding collagen origins and evolution is an essential piece to the puzzle of metazoan phylogeny and the evolution of multicellularity. For vertebrate chordates, including mammals, collagens are critical in formation of skin, bone, teeth, and other tissues. In humans collagens are the most abundant protein produced, making up 25% of body mass (Alberts et al., 1994). Many diseases are associated with mutations in collagens or breakdown of the collagen molecular pathway (Leitinger and Hohenester, 2006), and collagen pro-peptides may have anti-tumor and anti-angiogenic regulation properties (Richard-Blum and Ballut, 2011). Here, we report that *Pleurobrachia bachei*, a ctenophore, has 7 full length collagen type IV genes at 5 distinct genetic loci, as well as 4 additional collagen proteins which we were unable to assign to a specific family.

All collagens are secreted proteins, and are characterized by having three protein polymers wound into a diagnostic collagen triple helix. Each individual polymer in the triple helix is called an alpha chain, and a given collagen can be either a homotrimer (having three genetically similar alpha chains) or heterotrimer (a combination of two or three genetically distinct alpha chains). Though the triple helix is the most dominant feature of the collagen protein, non-collagen motifs surrounding the triple helix (pro-

peptides) are thought to be functionally important, playing a key role in the correct formation of the triple helix and in receptor-protein recognition (Leitinger and Hohenester, 2006). These polymers contain glycine at every third position, followed by two amino acids (X-Y). These triplets, called Gly-X-Y repeats, allow for the formation of a tightly wound triple helix, as Glycine is the smallest amino acid and thus allows for condensed packing formation. Collagens often contain over 1500 amino acids in their Gly-X-Y domains (Boot-Handford and Tuckwell, 2004). On the C terminal ends of collagen triple helixes are Non-collagen (NC1) domains called C-propeptides, while the N terminus contains a short helical triple helical domain (minor helix) and a second Non-collagen domain (NC2) (Boot-Hanford and Tuckwell, 2003). These terminal ends, C-propeptides and NC2, are insoluble and facilitate both the retention of the soluble inner triple helix collagen domain in solution and in the localization of collagen in tissues. These propeptides are cleaved as they exit the cell by specific enzymes (C- and N-proteinases), leaving the triple helix domain intact. Once extracellular and their propeptides are clipped, individual collagen helices will covalently bind to each other in a lateral formation, allowing for a variety of secondary structures including networks, microfilaments, and collagen fibrils. Collagen fibrils are the backbone of all epithelial tissues, crisscrossing the space between nearly every human cell, giving them structure and tensile strength (Alberts, et al. 1994).

There are a total of 29 different types of collagens described in humans, consisting of different combinations of the 48 described alpha chains. Collagens are named and numbered in the order they are discovered in vertebrate organisms, while there is some inconsistency in naming collagens in invertebrates. Types I, II, III, V, and

XI are called the fibrillar collagens, which are responsible for skeletogenesis in vertebrates. The two most abundant of the fibrillar collagens, types I and II, are responsible for proper formation of bone, teeth, and cartilage, respectively. Types III, V, and XI are less abundant than types I and II, but are essential for bone and cartilage tissues as well (Boot-Hanford and Tuckwell, 2003). Interestingly, fibrillar collagens are found across Metazoa, from Porifera and Cnidaria through Chordata. Few exceptions to this case are found, the occurrences being in the Arthropoda and Annelida. This Metazoa-wide collagen signature indicates that the metazoan common ancestor had the molecular toolkit for developing some sort of fibrillar collagen structures, even if they were rudimentary (Boot-Hanford and Tuckwell, 2003).

Surprising new evidence suggests that collagen IV may also be conserved across the metazoan lineage. Collagen IV is a key component of the Basal Lamina (BL), a layer of tissue which connects outer epithelial layers to tissues below. BL formation is considered by some to be indicative and a necessary hallmark of 'true epithelia' (Exposio et al., 2009). Until recently, collagen IV had been detected in Cnidaria, but in no other basal phyla. It has been shown that Cnidaria use collagen IV in development and healing (Fowler, et al. 2000), and possess a basal lamina-like membrane containing collagen IV (Shimizu, et al., 2007). A recent paper suggests that Porifera, one of the most basal phyla in Metazoa, have many of the molecular components necessary to make a primordial BL, including collagen type IV (Leys and Reisgo, 2011).

Collagen IV has recently been proposed to have evolved from sponging short-chain collagen (SSCC). SSCC's are defined as having 2 collagenous domains with 79 Gly-X-Y repeats and 3 NC domains. A hypothesis put forth by Aouacheria et al. 2006, is

that early multicellular animals first evolved SSCC in order to attach to some substrate in the marine environment, with collagen IV and BL formation evolving from co-adaptation of SSCC genetic pathways. Porifera still contain these SSCC's, and SSCC-related molecules are also present up through the non-vertebrate chordates. It is thought that the split between vertebrates and the rest of Metazoa was facilitated in some way by a duplication event within the SSCC family, allowing eventually the development of true collagen IV proteins. Phylogenetic comparison of SSCC to type IV NC domains to date support this theory, showing that NC1 domain in *E. mulleri* (a sponge in class Demospongiae) is homologous to the corresponding domain in type IV collagens, and that human ColVα1 and ColVα2 contain more than 75% homology with *P. jarrei* sponge SSCC domains (Aouacheria, et al, 2006). It has been shown that collagen molecules have evolved using standard gene duplication events, and this fact has been exploited to study evolution of associated HOX genes (Bailey et al., 1997). In some organisms, collagens are aligned in the genome in a pair-wise 'head-to-head' fashion along the chromosomes. That is to say that the NC2 domains are abutting, with one collagen coded in a forward direction and the other on the complementary strand in the reverse direction, indicating an inverted duplication in the ancestor. This inverted duplication has happened 3 times in vertebrate chordates, resulting in 7 alpha chains in type IV collagens (Hudson, et al. 1993). These 7 alpha chains are categorized into two sub-family types; α1-like and α2-like. In humans, the α1-like types are α1, α3, and α5, while the α2-like types are α2, α4, and α6. This 'type' patterning can be seen across the rest of Metazoa, where most of the more basal organisms have only one type IV α1-like and one α2-like chain (Figure 5/table 3).

Ctenophores, a group of jelly bodied animals that use ciliated comb rows for locomotion, are one of the most basal metazoans, as suggested by phylogenetics (Dunn, et al, 2008, Kohn, et al. 2011). Scientists argue as to their proper place on the universal animal tree, especially with respect to Porifera or Cnidaria. Some groups place Porifera as the most basal metazoan, mostly based on morphological evidence. Others argue that Ctenophores are the true metazoan ancestor, stating that their possession of the smallest mitochondrial genome yet found is a highly derived character state, indicating a pre-Porifera split (Dunn, et al, 2008, Kohn, et al. 2011). Recently, it has been shown that two separate sponges, namely *Sycon coactum* and *Corticum candelabrum*, have two distinct type IV collagens in each organism. These findings indicate that it is possible that α1 and α2 sub-family chains had diverged before Porifera, and that the previous 'SCCC to type IV' hypothesis may be incorrect (Leys and Riesgo, 2011). Therefore, the presence or absence of collagen IV or SSCC in ctenophores could shed light on their place in the metazoan tree, as well as give further information as to the evolution of epithelia and BL tissues. Furthermore, *P. bachei* may be a model organism to use for the study of collagen in human diseases, given their basal position. This paper focuses on collagen family types found in the *Pleurobrachia bachei* genome, recently sequenced at Friday Harbor Labs, in Washington state (Moroz, 2012, unpublished data). We show that *P. bachei* has 7 distinct full length α-chains (average = 2200 aa) of type IV collagen, each with multiple stretches of Gly-X-Y repeats (average = 13). Interestingly, two of the three collagen IV α1-like proteins found in *P. bachei* were aligned in a 'head-to-head' alignment typical of vertebrate chordates (Figure 3). We also show that 4 additional collagen-like proteins are present, though their collagen family associations are not yet known.

METHODS

Bioinformatics

We used 58 distinct *Homo sapiens* collagen alpha chain sequences, obtained from the Nucleotide database at National Center for Biotechnology Information (NCBI) (http://ncbi.nlm.nih.gov/) to find homologues in *P. bachei.* We then used TBlastN on the Moroz Lab gene database server, Neurobase, to obtain conserved protein sequences. With these sequences we ran BlastP against the NCBI database to gain conserved sequences from across Metazoa for gene tree comparisons. Some of the more basal organism sequences were obtained in the same way from the BROAD institute database (http://www.broadinstitute.org/) and from the Joint Genome Institute (JGI) database (http://www.jgi.doe.gov/). We used GeneDoc, MEGA, and ClustalX software to align our sequences. To obtain gene trees, we used MEGA tree software ML. Bootstrap support of less than 50 were omitted. We trimmed our sequences to the same region as described in Aouacheria et al. 2006, in order to get the most accurate phylogenetic trees, because Gly-X-Y repeats are highly divergent across all collagen families. To obtain intron/exon information, as well as general protein family domains, we used the Simple Modular Architecture Research Tool (SMART) server (http://smart.emBL-heidelberg.de/smart/set_mode.cgi?NORMAL=1).

Primers were designed using complementary DNA sequences for collagen IV αA-D. PCR products were cloned using OneShot cells grown on LB/KAN culture plates. Clone vectors were isolated with Qiagen Spin Mini-prep Kits, and sequenced by

SeqWright Genomic. Please contact me for primer sequences and full collagen sequences.

Animals for in situ hybridizations were collected off the docks at Friday Harbor Labs in San Juan Island, Washington between the months of April through August 2012. Animals were kept in circulating salt water tanks for up to two weeks prior to fixation.

In situ protocol

Whole animals were fixed in 4% paraformaldehyde in Filtered Sea Water (FSW) over night at 4^{0}C. Next we did three ten minute washes in PTW (PBST) at room temperature, followed by a 1:1 10 minute wash in Methanol/PTW at room temperature in order to equilibrate the animal to MeOH. We then stored our specimens in 100% MeOH at -20^{0}C for at least two hours, or up to two weeks. We then rehydrated the specimens with a succession of 10 minute washes in MeOH/PTW (3:1, 1:1, 1:3, 0:1) at room temperature. Next we washed our samples in a 1:1 solution of pre-hybridization buffer (pHB) and PTW for 15 minutes at room temperature. We then incubated our samples for an hour at 60^{0}C in pure pHB, followed by hybridization in hybridization buffer (HB) with 1-5 μL of DIG-RNA probe over night at 60^{0}C. Subsequently we washed our specimens in HB for 30 minutes at 60^{0}C, followed directly with a 1:1 wash in HB/PTW for 30 minutes at room temperature, and then in PTW for 30 minutes at room temperature. We blocked hybridization using a 10% goat serum (GS) for 60 minutes at room temperature, then incubated over night at 4^{0}C in a 1/2000 dilution of anti-DIG. Finally we performed four 30 minute washes in PBS at room temperature. We then would aliquot 1 mL of detection buffer into clean 5 mL wells for each specimen to develop. When we were

ready to develop, we added 20 µL of NBT/BICP and mixed to dissolve, then added one sample per well and covered and placed on ice. Development of in situ expression would range in time from 1 hour to approximately 6 hours, which we would stop by washing in 4% paraformaldehyde in MeOH for 5 minutes, then storing in pure Ethanol (EtOH) for up to a week. To mount in situs, we would dissect desired areas of the Ctenophores in EtOH, then add them into a new well in Methysalicylate until they sank. We then placed animals onto glass slides and fixed in place using Permount. Photos of in situs were taken with a Nikon compound microscope and camera.

RESULTS

We used 58 distinct *Homo sapiens* collagen sequences (*COLIα1, COLIα2, COLIα3,* etc.) obtained from NCBI to pull up conserved proteins in *Pleurobrachia bachei* using the Moroz lab genome server Neurobase. We found 7 distinct homologues to type IV collagen α-chains, and 4 non-full length sequences, each with conserved collagen Gly-X-Y repeats and other collagen protein family markers. The 4 remaining non-type IV were unclassifiable because they were not full length, or did not otherwise contain the pro-peptides that allows for genetic relationship inference.

We did not find any SSCC homologues in the *P. bachei* genome, only full length type IV collagen homologues, which we dubbed homologues *PbalphaA-G* (*PbalphaA:* 2272aa, *PbalphaB:* 2448aa, *PbalphaC:* 2021aa, *PbalphaD:* 2395aa, *PbalphaE:* 1753aa, *PbalphaF:* 2402aa, *PbalphaG: 1459*aa). Phylogenetic analysis showed that of the type IV α-chains, there were two groupings of proteins, similar to that of α1-like and α2-like collagen type IV sub-families in *H. sapiens.* We therefore divided our collagen sequences

into two sub-families, the α1-like which contained *PbalphaA-C* and α2-like which contained *PbalphaD-F* (time constraints did not allow us to include *PbalphaG* in our gene tree analysis by time of publication). Interestingly, the sub-family collagen chains found in *P. bachei* are more diverse when compared intra-genetically than *H sapiens* sub-family collagen chains (Figure 5). Only the α2-like chains had repeat domains (RPT) in *P. bachei,* which may be important in their sub-family separation. Our gene alignment showed that the 12 cysteine residues found in NC1 domains were conserved from the *H. sapiens* sequence in *P. bachei* (Figure 5). We also note considerable homology in the β-hairpin regions of the NC1 domains. SMART protein domain analysis showed that each of these full length sequences contained a minimum of 7 Gly-X-Y regions, and a maximum of 21. Each homologue also contained two NC1 domain sequences conserved to specific collagen IV proteins, while only one recovered NC2 (N-terminal) pro-peptides. (Figures 1 and 2).

We found that two of our type IV α1-like proteins, *PbalphaB* and *PbalphaC,* were aligned on our scaffold in a pair-wise head-to-head fashion, similar to the arrangement found in the fibrillar collagens of some sponges and vertebrate chordates (Figure 3) c Furthermore, we found that an apparent in-line gene duplication (traditional gene duplication) event happened with *P. bachei's* α1-like *PbalphaA* and *PbalphaB,* as well as with α2-like *PbalphaD* and *PbalphaE.* We deduced this as they are directly in line along the scaffold. Lastly, we found two collagen IV α2-like homologues on two separate scaffolds, *PbalphaF* and *PbalphaG.* Our *PbalphaF* α2-like sequence ran nearly the entire length of our scaffold, which raises the possibility that there are more α2-like in *P.*

bachei if the in-line duplication pattern was present in this particular gene also. We have insufficient data to prove this, however.

Preliminary in situ hybridization of *PbalphaC* shows diverse expression across tissues, most notably at the base of the comb rows, along the underlying meridional canals, in the tissues holding the statolith, and in the dermal tissues. Single cell expression was also notable. These results have not yet been duplicated, however, and further in situs are needed to confirm the location of collagen IV in *P. bachei*.

DISCUSSION

We found a total of 7 distinct type IV collagens in *Pleurobrachia bachei,* forming two sub-families in genetic analysis which we called α1-like chains (*PbalphaA-C*) and α2-like chains (*PbalphaD-F).* It has recently been suggested the sponge *Pseudocorticium jarrei* has type IV collagens, with new evidence unveiling two distinct collagen IV sequences in *Sycon coactum* and *Corticum candelabrum*. These findings indicate that the evolution of α1-like and α2-like type IV collagen sub-families found in derived organisms could have predated the sponge ancestor (Leys and Riesgo, 2011). Our findings lend support to this theory, because of the distinct separation in *P. bachei* type IV collagens into α1-like and α2-like sub-families, and the divergence of sequences within each sub-family's α-chains. Our results do not support the 'SCCC to type IV' theory proposed by Aouacheria et al. 2006. The separation of α1-like and α2-like sub-families in ctenophores suggests instead that collagens had evolved these genetic divergences before the split of Ctenophora and the rest of Metazoa. Human collagen type

IV profiles did not show significant bootstrap support to indicate divergence between α1 and α2 sub-family α-chains, as was found in *P. bachei.*

Furthermore, *P. bachei* has the most extensive expansion of type IV collagens of any animal yet described, with 7 complete sequences. The only organisms that come close to this count are the vertebrate chordates, with 6 type IV α-chains (figure 5/table 3). These facts show that *P. bachei* has a more diverse type IV collagen profile than vertebrate chordates, including *H. sapiens.* This makes sense when we consider the amount of time since the divergence of Ctenophora and bilateria. There are two hypotheses that are equally plausible to explain this expansion. One, which says ctenophores have always had a diverse collagen profile, and an extensive loss of genetic variation was seen at the time of their divergence from Metazoa. And two, which says that the ctenophores had a primitive collagen profile at the time of their divergence, and convergent evolution explains their extensive collagen expansion. We may never know which is the correct hypothesis, but either option is extremely interesting.

The fact that a pair of collagens were found in a pair-wise head-to-head fashion (*PbalphaB, PbalphaC*), and 5/7 were found to be associated with some sort of gene duplication event along the *P. bachei* genome, is another significant attribute of *P. bachei's* collagen profile. The head-to-head organization of *P. bachei's* collagens, seen also in some sponges and the vertebrate chordates, indicates that there is some conserved functional relationship between inverted gene duplication events and collagen proteins. What the exact functional relationship is is not yet known.

Preliminary in situ hybridizations of *PbalphaC* revealed expression patterns in a variety of tissues. This is what we would predict considering the function of type IV collagen proteins in the rest of metazoa as a key molecule in BL tissues (Figure 5). Most interestingly is the even expression patterns in single cells in the dermal layers of the Ctenophore. The diverse nature of expression of collagen in *P. bachei*, and their close relationship to the Porifera phyla, again suggest either an extreme expansion and use of collagen in Ctenophora, or an extreme reduction in sponges.

It is interesting to note that we were unable to find any fibrillar collagen family proteins in *P. bachei*, considering that it has been found across all Metazoa to date (Boot-Hanford and Tuckwell, 2003). We did discover 4 additional collagen proteins in *P. bachei* with conserved diagnostic Gly-X-Y repeats, but these were not full length and none contained either C-terminal or N-terminal pro-peptides. These 4 additional collagens could be the fibrillar collagens we seek, but at this point in time our results remain inconclusive. One additional hypothesis we put forth is the idea that ctenophores could have adapted the type IV collagen to serve the same functions as fibrillar collagen. This could be an intriguing area for future research.

CONCLUSIONS

We found 7 distinct collagen type IV α-chains, 5 of which were associated with gene duplication events, as well as 4 non classifiable collagen sequences in *P. bachei.* This type of expansion in collagen profiles has not yet been seen in any other organism basal organism, and the extreme expansion of type IV collagens hasn't been seen in any

organism to date, including the vertebrate chordates, which have only 6 type IV collagen α-chains. The sub-family groupings into α1-like and α2-like observed in *P. bachei* support the hypothesis that this genetic divergence happened before the split of Ctenophora and Bilateria, indicating that the "SCCC to type IV' hypothesis may be incorrect. Gene duplication events, including 'head-to-head' duplication, further bolster the diverse nature of this ctenophore's type IV collagen profile, and suggest a conserved functionality between inverted gene duplication events and collagen proteins. Expression across multiple cell types observed during in situ experiments again indicates that this organism is using collagen to a greater extent than has previously been observed in basal organisms. Our results combined suggest that the ancestor to the metazoan tree had a more derived and diverse collagen profile than previously thought, and an extreme loss of genetic diversity has occurred in the Porifera phyla. Alternatively, one might interpret our results as suggestive of extensive convergent evolution between the collagen profiles of vertebrate chordates and *Pleurobrachia bachei.*

TABLES & FIGURES

Collagen IV α1-like chains in P. bachei

Pbalpha-A	Begin	End	*Pbalpha-B*	Begin	End	*Pbalpha-C*	Begin	End
Gly-X-Y	1	57	Gly-X-Y	332	395	C-SP	1	19
Gly-X-Y	47	112	Gly-X-Y	381	452	Gly-X-Y	64	117
Gly-X-Y	97	165	Gly-X-Y	439	498	Gly-X-Y	90	164
Gly-X-Y	203	273	Gly-X-Y	496	552	Gly-X-Y	141	202
Gly-X-Y	392	455	Gly-X-Y	551	610	Gly-X-Y	264	323
Gly-X-Y	442	514	Gly-X-Y	817	876	Gly-X-Y	614	673
Gly-X-Y	500	558	Gly-X-Y	921	986	Gly-X-Y	674	731
Gly-X-Y	557	612	Gly-X-Y	959	1026	Gly-X-Y	728	784
Gly-X-Y	611	668	Gly-X-Y	1080	1140	Gly-X-Y	750	808
Gly-X-Y	653	721	Gly-X-Y	1136	1188	Gly-X-Y	875	934
Gly-X-Y	699	775	Gly-X-Y	1186	1241	Gly-X-Y	1070	1142
Gly-X-Y	759	828	Gly-X-Y	1576	1636	Gly-X-Y	1114	1175
Gly-X-Y	812	876	Gly-X-Y	1708	1749	Gly-X-Y	1174	1231
Gly-X-Y	918	975	Gly-X-Y	1756	1819	Gly-X-Y	1223	1289
Gly-X-Y	974	1035	NC1	1822	1944	Gly-X-Y	1264	1343
Gly-X-Y	1182	1248	NC1	1982	2071	Gly-X-Y	1319	1380
Gly-X-Y	1241	1301				Gly-X-Y	1573	1637
Gly-X-Y	1300	1365				Gly-X-Y	1675	1743
Gly-X-Y	1448	1515				Gly-X-Y	1718	1787
Gly-X-Y	1630	1696				Gly-X-Y	1774	1837
Gly-X-Y	1684	1745				Gly-X-Y	1836	1897
Gly-X-Y	1833	1890				Gly-X-Y	1871	1949
Gly-X-Y	1887	1937				NC1	2049	2172
Gly-X-Y	1936	1992				NC1	2173	2270
Gly-X-Y	2118	2178						
Gly-X-Y	2178	2244						
NC1	2249	2367						
NC1	2368	2444						

Table 1: **Above are the conserved protein domains in α1-like chains (*Pbalpha-A, Pbalpha-B, Pbalpha*-C) found in *Pleurobrachia bachei* gene models, recently sequenced at the Moroz lab at the University of Florida. 'Begin' and 'End' columns show the amino acid position along the length of the protein. Most notable are the Gly-X-Y protein families, which are diagnostic of true collagen alpha chains and have only recently been found in other basal Metazoans. Each α1-like chain also contains two conserved NC1 domains, highlighted in red, which we later used to deduce gene trees and infer relationships across metazoa. One C-terminal signal peptide domain was found in *Pbalpha-C*, marked as C-SP. These tables were constructed using the Simple Modular Architecture Research Tool (SMART) server (http://smart.emBM-heidelberg.de/smart/set_mode.cgi?NORMAL=1). Total amino acid counts: *Pbalpha-A: 2448, Pbalpa-B: 2021, Pbalpha-C: 2272.***

Collagen IV α2-like chains in P. bachei

PbalphaD	Begin	End
Gly-X-Y	62	126
Gly-X-Y	85	160
Gly-X-Y	234	290
Gly-X-Y	289	352
Gly-X-Y	341	409
Gly-X-Y	942	1003
Gly-X-Y	1049	1113
Gly-X-Y	1109	1170
Gly-X-Y	1156	1214
Gly-X-Y	1535	1592
Gly-X-Y	1581	1639
Gly-X-Y	1639	1697
Gly-X-Y	1720	1778
Gly-X-Y	1826	1883
Gly-X-Y	1855	1908
Gly-X-Y	1907	1965
Gly-X-Y	1953	2015
Gly-X-Y	2041	2100
Gly-X-Y	2092	2152
NC1-D	2163	2273
NC1-D	2275	2391

PbalphaE	Begin	End
Gly-X-Y	1	59
Gly-X-Y	140	209
Gly-X-Y	200	264
Gly-X-Y	241	314
Gly-X-Y	529	603
Gly-X-Y	837	914
Gly-X-Y	891	963
Gly-X-Y	945	1008
Gly-X-Y	1007	1066
Gly-X-Y	1040	1123
Gly-X-Y	1175	1239
Gly-X-Y	1226	1290
Gly-X-Y	1282	1343
Gly-X-Y	1372	1433
Gly-X-Y	1425	1486
NC1-D	1495	1605
NC1-D	1606	1722

PbalphaF	Begin	End
C-SP	1	26
Gly-X-Y	178	234
Gly-X-Y	220	279
Gly-X-Y	278	337
Gly-X-Y	324	389
Gly-X-Y	422	485
Gly-X-Y	475	540
Gly-X-Y	535	595
Gly-X-Y	1078	1151
Gly-X-Y	1422	1486
Gly-X-Y	1479	1542
Gly-X-Y	1529	1587
Gly-X-Y	1743	1818
Gly-X-Y	1950	2008
Gly-X-Y	1994	2057
NC1-D	2137	2248
NC1-D	2249	2364

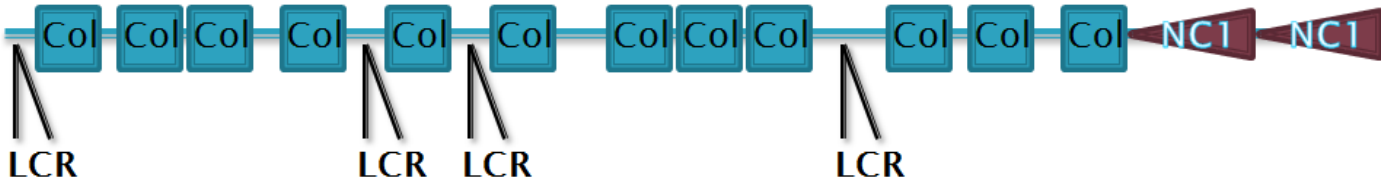
PbalphaG	Begin	End
Gly-X-Y	40	102
Gly-X-Y	82	157
Gly-X-Y	139	208
Gly-X-Y	270	324
Gly-X-Y	406	475
Gly-X-Y	452	515
Gly-X-Y	542	603
NC1-D	665	775
NC1-D	776	872
NC1-D	901	1011
NC1-D	1012	1109

Table 2: **Above are the conserved protein domains in α2-like chains (*Pbalpha-D, Pbalpha-E, Pbalpha*-F, *Pbalpha*-F) found in *Pleurobrachia bachei* gene models, recently sequenced at the Moroz lab at the University of Florida. 'Begin' and 'End' columns show the amino acid position along the length of the protein. Most notable are the Gly-X-Y protein families, which are diagnostic of true collagen alpha chains and have only recently been found in other basal Metazoans. Each α1-like chain also contains two conserved NC1 domains (NC1-D), highlighted in red, which we later used to deduce gene trees and infer relationships across metazoa. One C-terminal signal peptide domain was found in *Pbalpha-F*, marked as C-SP. These tables were constructed using the Simple Modular Architecture Research Tool (SMART) server (http://smart.emBM-heidelberg.de/smart/set_mode.cgi?NORMAL=1). Total amino acid counts: *Pbalpha-D: 2395, Pbalpa-E: 1753, Pbalpha-F: 2402.***

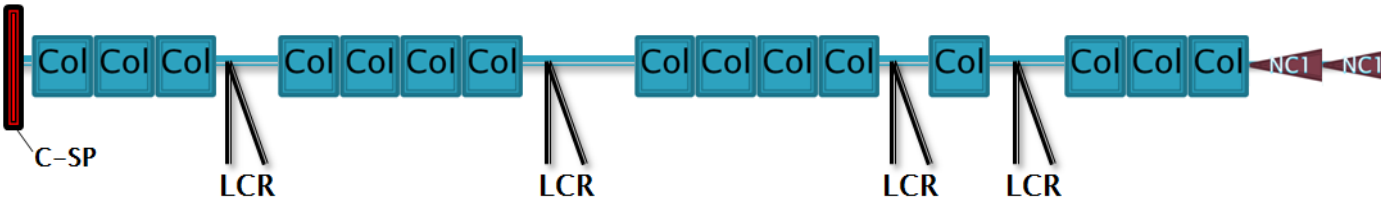
α1-Like chain motifs in P. bachei

Pbalpha-A (2448 amino acids)

Pbalpha-B (2021 amino acids)



Pbalpha-C (2272 amino acids)

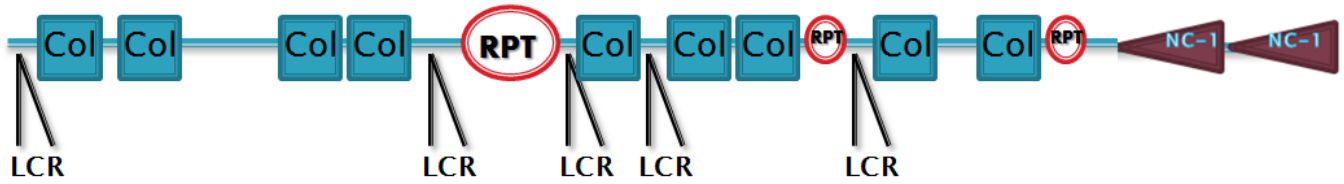


Symbol	Meaning
COL	Gly-X-Y domain
NC1	NC1 domain
LCR	Low Complexity Region
C-SP	C-terminal Signal Peptide

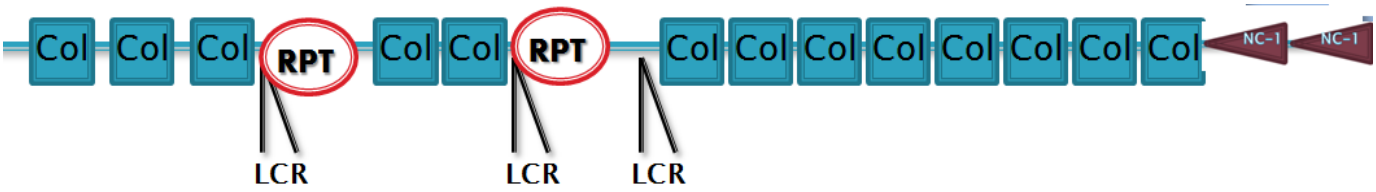
Figure 1: **Collagen IV α1-Like chain motifs found in *P. bachei*. Note the large amount of collagen Gly-X-Y motifs in each gene (*Pbalpha-A: 21 Gly-X-Y domains, Pbalpha-B: 12 Gly-X-Y domains, Pbalpha-C: 15 Gly-X-Y domains).* Each gene also contained two NC-1 domains, and several Low Complexity Regions (LCR). These LCR are places of high amino acid repeats or scaffold gaps in sequence. All figures were generated with information from the Simple Modular Architecture Research Tool (SMART) server (http://smart.emBM-heidelberg.de/smart/set_mode.cgi?NORMAL=1).**

α2-Like chain motifs in P. bachei

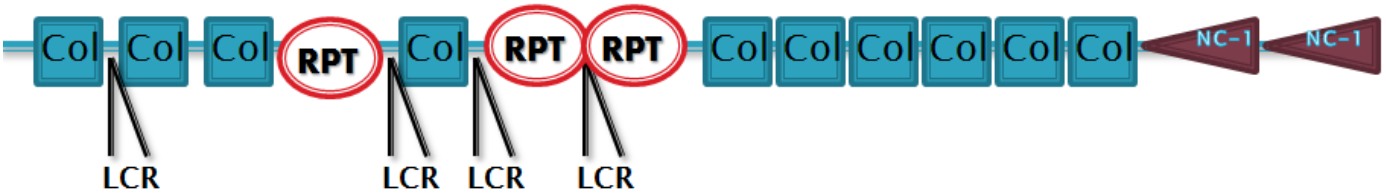
Pbalpha-F (2402 amino acids)



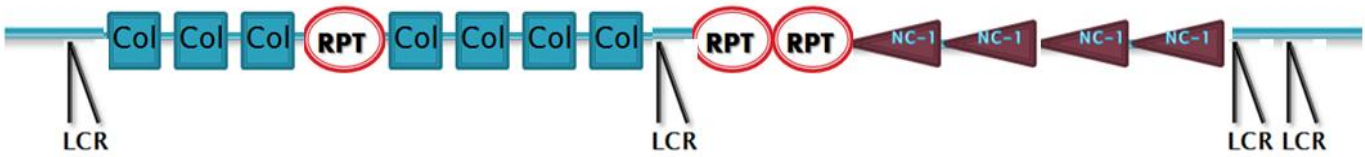
Pbalpha-D (2395 amino acids)



Pbalpha-E (1753 amino acids)



Pbalpha-G (1459 amino acids)



Symbol	Meaning
COL	Gly-X-Y domain
NC1	NC1 domain
RPT	Internal repeat
LCR	Low Complexity Region
C-SP	C-terminal Signal Peptide

Figure 2 (previous page): **Collagen IV α1-Like chain motifs found in *P. bachei*. Note the large amount of collagen Gly-X-Y motifs in each gene (*Pbalpha-A: 21 Gly-X-Y domains, Pbalpha-B: 12 Gly-X-Y domains, Pbalpha-C: 15 Gly-X-Y domains)*. Each gene also contained two NC-1 domains, and several Low Complexity Regions (LCR). These LCR are places of high amino acid repeats or scaffold gaps in sequence. All figures were generated with information from the Simple Modular Architecture Research Tool (SMART) server (http://smart.emBM-heidelberg.de/smart/set_mode.cgi?NORMAL=1).**

Scaffold organization in type IV alpha chains

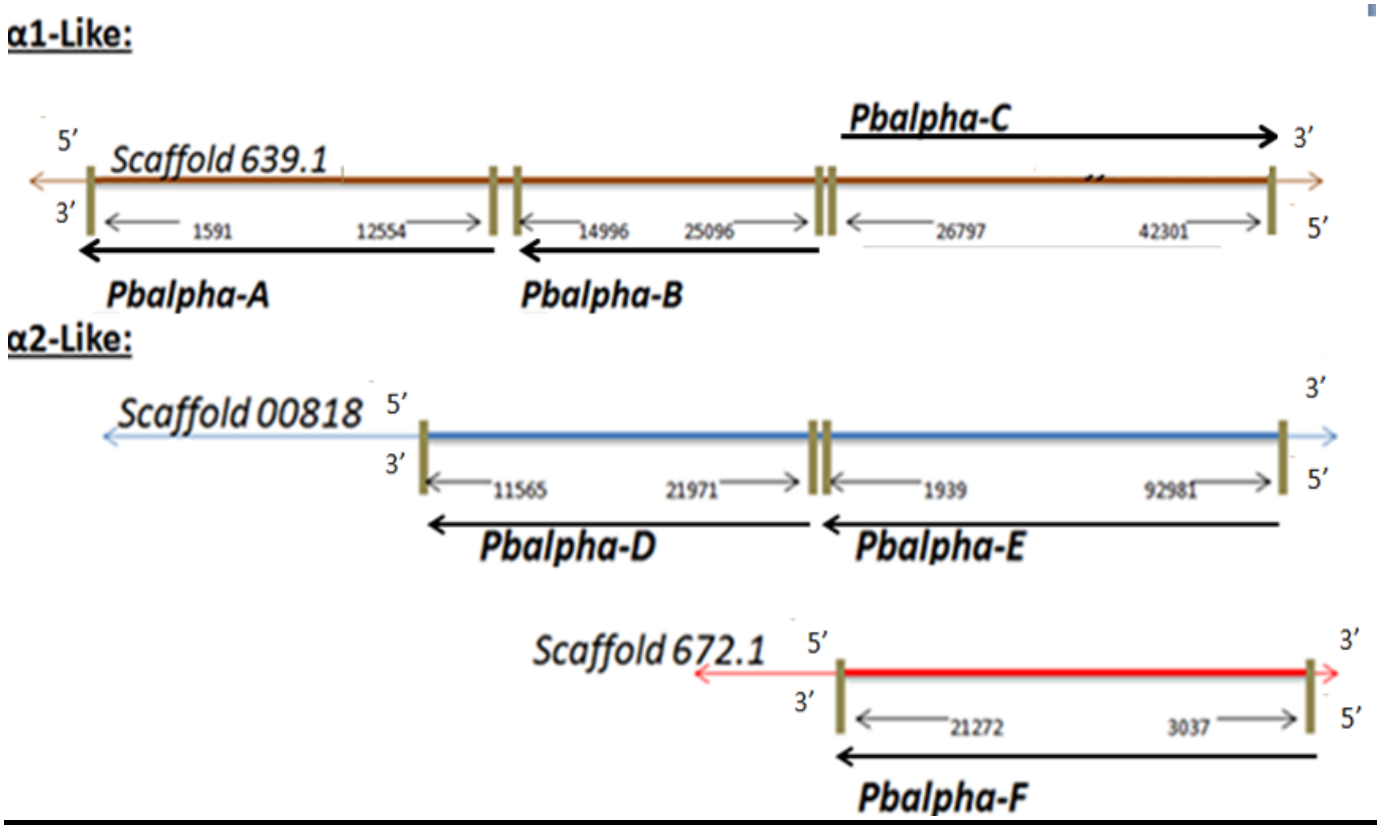


Figure 3: **Scaffold alignment in *Pleurobrachia bachei* alpha chains. Thick colored shading indicates protein coding region along scaffold, while thinner like-colored arrows indicate entire scaffolds. Thick black arrows beneath gene names (*Pbalpa-A, Pbalpha-B* etc.) indicate directionality along scaffold. Small numbers with arrows pointing to vertical tan bars below the scaffolds represent amino acid start and stop positions. Notice the 'head-to-head' alignment between *Pbalpha-B* and *Pbalpha-C,* indicating an inverted gene duplication. Note also the traditional gene duplications indicated by position along scaffold between *Pbalpha-A&B* and *Pbalpha-D&E*. Figure generated using the Moroz lab *Pleurobrachia bachei* gene model server.**

Type IV Collagen and Spongin Short Chain Collagen Gene Alignments

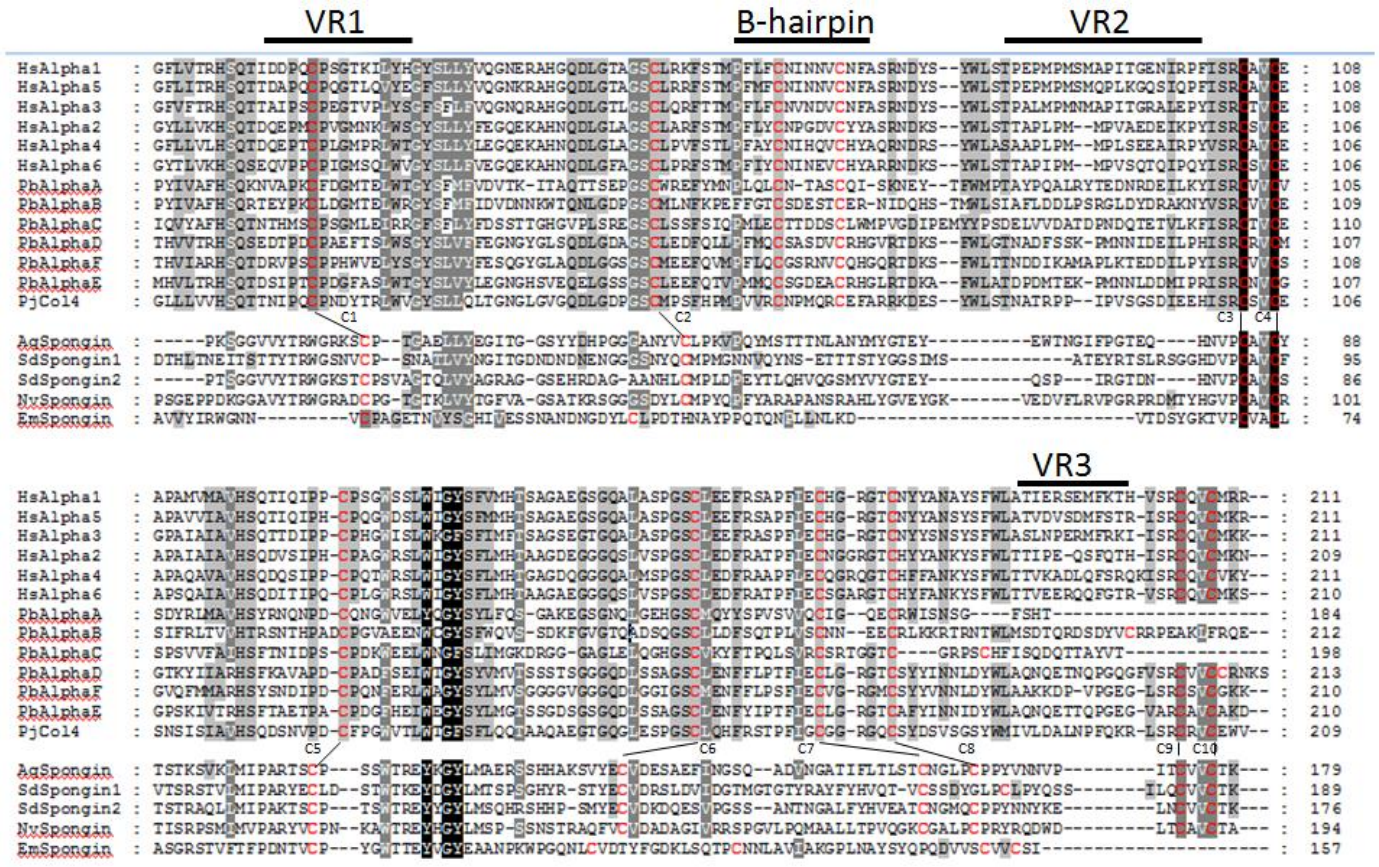


Figure 4: Multiple species alignments of Spongin Short-Chain Collagens (SSCC) and collagen Type IV alpha chain NC1 domains. Our alignments were produced using GeneDoc software. (Gene names and related organisms: Hsapha1-6: *Homo sapiens*, PbalphaA-E: *Pleurobrachia bachei*, PjCol4: *Pseudocorticum jarrei*, Agspongin: *Anopheles gambiae*, Sdspongin1-2: *Suberites domuncula*, Nvspongin: *Nematostella* vectensis, Emspongin: *Ephydatia mulleri*). Highlighted in red are Cystein domains, marked C1-10, conserved across SSCC and type IV alpha chains. Black shading indicates 100% conservation across phyla, while grey shading indicates partial homology. Noted above the alignments are NC1 specific domains VR1-3 (variable regions), and β-hairpin region. Homology between *Pleurobrachia bachei* and other type IV collagens is readily apparent.

In situ expression of PbalphaC and collagen type IV gene tree

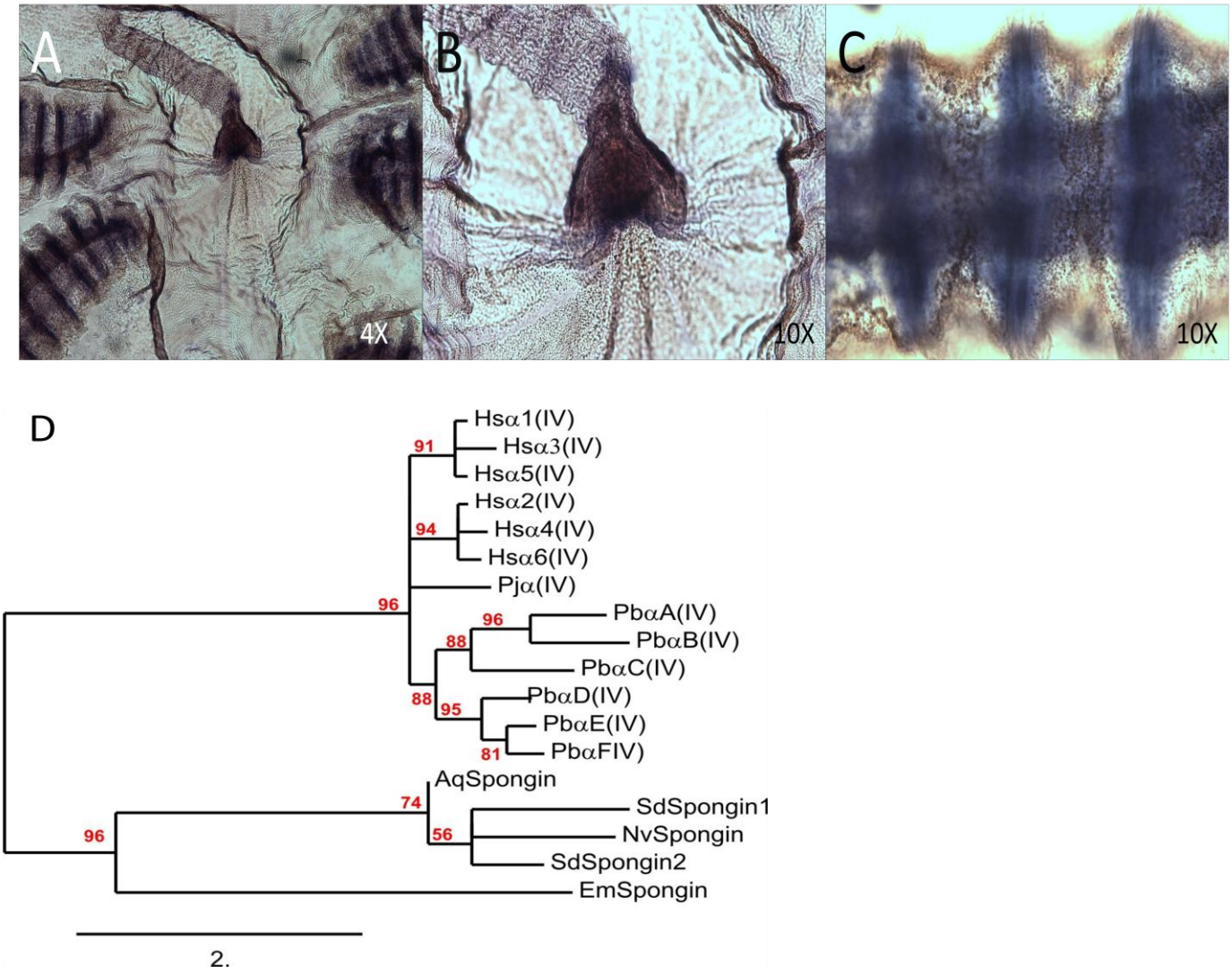
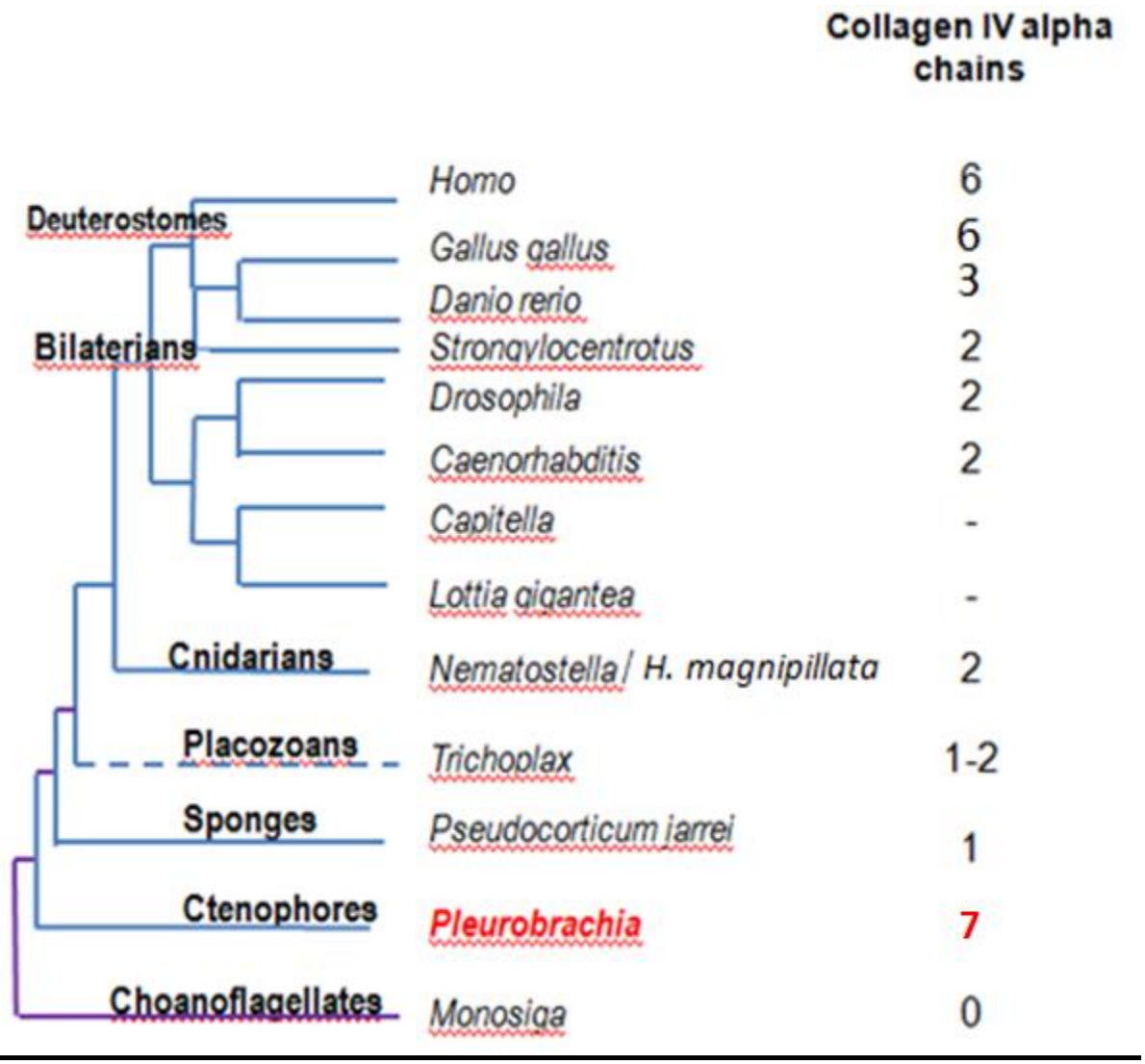


Figure 5: A-C show in situ preparations (combs oriented into the page) using *PbalphaC* . A: Note expression along the base of the combs, along the meridional canal (above the comb rows), and in the statolith (center). B: Close examination of the statolith reveals heavy expression in tissues composing the endoderm of the statolith, and specific single cell expression in the surrounding dermal layers. C: 10X photo of comb rows showing expression of *PbalphaC* in single cells surrounding comb rows, at the base of the comb rows, and along the meridional canals. D: Gene tree showing alignment of *P. bachei* collagen type IV alpha chains(Hsapha1-6: *Homo sapiens*, PbalphaA-E: *Pleurobrachia bachei*, PjCol4: *Pseudocorticum jarrei*, Agspongin: *Anopheles gambiae*, Sdspongin1-2: *Suberites domuncula*, Nvspongin: *Nematostella* vectensis, Emspongin: *Ephydatia mulleri*).. Note the two part grouping of *PbαA-C* and *PbαD-F*. Notice also that alpha chains in *P. bachei* show a more diverse intra-gene relationship than human alpha chains. Figure generated using MEGA tree software ML, bootstrap support of less than 50 omitted.
Probe Sequence on next page.

ATCTGGTATGTTGGAGATTAGGCGAGGTTTCTCCTTCCTCTACTTTGACTCTTCAACCGGTACTGGTCACGGTGTCC
CCCTGAGCAGAGAGGGAAGTTGCCTTTCATCTTTCTCCATCCAAC
CTATGCTGGAGTGTACTACCAAATACTGCGAGGTAAAGGGAGACGACAGCTCTCTCTGGATGCCAGTGGGAGATA
TTCCAGAGGGCAGTGCCGCTGGTATGTATTACCCTTCGGATGAACT
AGTGGTTGACGCGACAGATCCTAATGATCAAACTGAGACTGTCCTCAAGTTCATCTCTCGATGTACGGTCTGTGAA
TCACCATCAGTTGTGTTTGCCATTCACTCCTTCACCAACATTGACCC
CTCATGTCCTGATAAATGGGAGGAGCTATGGAATGGATTCTCTCTCATCATGGGTAAGGATCGAGGTGGTGGTGCT
GGGCTGGAACTACAGGGTCACGGATCTTGTGTCAAATACTTCACTCC
TCAGCTTTCCGTACGGTGCAGCCGAACCGGAGGAACAACTGGACGACCGAGAAGGGCGAATTCGTTTAAACCTGC
AGGACTAGTCCCTTTAGTGAGGGTTAATTCTGAGCTTGGCGTAATCA
TGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATAAAGT
GTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTG

Type IV Collagen alpha chains across Metazoa



PHYLA	CLASS	Species	COLLAGEN 4 COUNT	ACCESSION #
Porifera	Homoscleromorpha	*Pseudocorticum jarrei*	1	Q7JMZ8
Ctenophora	Tentacula	*Pleurobrachia bachei*	7	
Cnidaria	Hydrozoa	*H. magnipillata*	2	DN138061, DN636820
	Anthozoa	*Nematostella vectensis*	2	c438003055.Contig1, c415700716.Contig2
Arthrodpoda	Insecta	*D. melanogaster*	2	O18407, P08120
		Bombyx mori	2	BP125709, CK522520
		Anopheles gambiae	2	Q7PVR8, BM588632
Nematoda	Chromadorea	*C. elegans*	2	P1740, P17139
Echinodermata	Echinoidea	*S. pupuratus*	2	Q26640, Q07265
Placozoa	tricoplacia	*Trichoplax adhaerens*	1 or 2	XP_002116198.1, XP_002116198.1
Chordata	Ascidiacea	*C. intestinalis*	2	BW439169, BW229076
	Actinopterygii	*D. rerio*	3	
	Cephalochordata	*B. floridae*	2	BW844807, BW895761
	Aves	*Gallus gallus*	6	Q919K3, XP_416952, AAY43819, XP_422615, XP_420320, XP_420322
	Mammalia	*M Musculus*	6	NM_009931, NM_009932, NM_007734, NM_007735, NM_007736, NM_053185
	Mammalia	*H. sapiens*	6	P02462, P08572, Q01955, P53420, P29400, Q14031

Figure 5/Table 3: Tree shows collagen type IV alpha chain count across the metazoan tree. The most notable data in the tree is the fact that *P. bachei* has 6 distinct alpha chains, similar to the collagen profile of *H. sapiens.* This is again noticeable in table 3, which shows that *P. bachei* has a collagen count more similar to that of vertebrate chordates than other basal metazoans. This figure and table were adapted from Aouacheria et al., 2011.

FIBRILLAR COLLAGENS ACROSS METAZOA

Phylum/Gene	Species/NCBI accession	Phylum/Gene	Species/NCBI accession
Freshwater Sponge	E. mulleri	Abalone	(Haliotis discus)
Emu1α	P18856, Q06452	Hdcol1α	O97405
		Hdicol2α	O97406
Sponge	(A. queenslandia)	Sea Urchin	S. purpuratus
		Spu1α	Q26634
Amq1α	WGS	Spu2α	Q26639
Amq2α	WGS	Spu6α	XP_001192332
Amq3α	WGS		
Amq4α	WGS	Sea Urchin	Paracentrotus lividus
Amq5α	WGS	Pli5α	CAE53096
Amq6α	WGS		
Amq7α	WGS	Ascidian	(Ciona intestinalis)
		Cin759	ci0100150759
Hydra	(H. Magnipapillata)	Cin301	ci0100154301
HCol1	WGS	Cin606	ci0100131606
HCol2	WGS	Cin916	ci0100144916
HCol3	WGS		
HCol5	WGS	Human	Homo sapiens
		Hsaα1(I)	P02452
Sea anenome	(N. vectensis)	Hsaα2(I)	P08123
Nve1α	ID # 25350	Hsaα1(II)	PQ14027
Nve2α	ID #6496	Hsaα1(III)	P02461
Nve3α	ID #21796	Hsaα1(V)	P20908
Nve4α	ID #36007	Hsaα2(V)	P05997
Nve5α	ID #1737	Hsaα3(V)	P25940
Nve6α	ID #26644	Hsaα1(XI)	P12107
Nve7α	ID #166	Hsaα2(XI)	P13942
Nve8α	ID #40962	Hsaα1(XXIV)	Q7Z5L5
		Hsaα1(XXVII)	Q8IZC6
Mosquito	(Anopheles gambiae)		
Agaα1	ENSANGG00000016690		
Agaα2	ENSANGG00000018512		
Honeybee	Apis mellifera		
Ame1α	AADG02012535		
Ame2α	AADG02005865		

(From Exposito et al, 2008)

References

Aouacheria, A., C. Geourjon, et al. (2006). "Insights into early extracellular matrix evolution: spongin short chain collagen-related proteins are homologous to basement membrane type IV collagens and form a novel family widely distributed in invertebrates." Mol Biol Evol 23(12): 2288-2302.

Bruce Alberts, D. B., Julian Lewis, Martin Raff, Kieth Roberts, James D. Watson (1994). Molecular Biology of the Cell, Garland Publishing, Inc.

Bailey, W. J., J. Kim, et al. (1997). "Phylogenetic reconstruction of vertebrate Hox cluster duplications." Mol Biol Evol 14(8): 843-853.

Boot-Handford, R. P. and D. S. Tuckwell (2003). "Fibrillar collagen: the key to vertebrate evolution? A tale of molecular incest." Bioessays 25(2): 142-151.

Dunn et al., (2008). "Broad phylogenomic sampling improves resolution of the animal tree of life." Nature, 452 (2008), pp. 745–749.

Exposito, J. Y., C. Larroux, et al. (2008). "Demosponge and sea anemone fibrillar collagen diversity reveals the early emergence of A/C clades and the maintenance of the modular structure of type V/XI collagens from sponge to human." J Biol Chem 283(42): 28226-28235.

Exposito, J. Y., U. Valcourt, et al. (2010). "The fibrillar collagen family." Int J Mol Sci 11(2): 407-426.

Fowler, S. J., S. Jose, et al. (2000). "Characterization of hydra type IV collagen. Type IV collagen is essential for head regeneration and its expression is up-regulated upon exposure to glucose." J Biol Chem 275(50): 39589-39599.

Hudson, B. G., S. T. Reeders, et al. (1993). "Type IV collagen: structure, gene organization, and role in human diseases. Molecular basis of Goodpasture and Alport syndromes and diffuse leiomyomatosis." J Biol Chem 268(35): 26033-26036.

Kohn, A. B., M. R. Citarella, et al. (2012). "Rapid evolution of the compact and unusual mitochondrial genome in the ctenophore, Pleurobrachia bachei." Mol Phylogenet Evol 63(1): 203-207.

Leitinger, B. and E. Hohenester (2007). "Mammalian collagen receptors." Matrix Biol 26(3): 146-155.

Leys, S. P. and A. Riesgo (2011). "Epithelia, an evolutionary novelty of metazoans." J Exp Zool B Mol Dev Evol.

Ricard-Blum, S. and L. Ballut (2011). "Matricryptins derived from collagens and proteoglycans." Front Biosci 16: 674-697.

Shimizu, H., R. Aufschnaiter, et al. (2008). "The extracellular matrix of hydra is a porous sheet and contains type IV collagen." Zoology (Jena) 111(5): 410-418.

Tyler, S. (2003). "Epithelium--the primary building block for metazoan complexity." Integr Comp Biol 43(1): 55-63.